

The distribution of the quasispecies for the Wright–Fisher model on the sharp peak landscape

Joseba Dalmau
Université Paris Sud and ENS Paris

March 28, 2014

Abstract

We consider the classical Wright–Fisher model with mutation and selection. Mutations occur independently in each locus, and selection is performed according to the sharp peak landscape. In the asymptotic regime studied in [3], a quasispecies is formed. We find explicitly the distribution of this quasispecies, which turns out to be the same distribution as for the Moran model.

1 Introduction

The concept of quasispecies first appeared in 1971, in Manfred Eigen’s celebrated paper [8]. Eigen studied the evolution of a population of macromolecules, subject to both selection and mutation effects. The selection mechanism is coded in a fitness landscape; while many interesting landscapes might be considered, some have been given more attention than others. One of the most studied landscapes is the sharp peak landscape: one particular sequence—the master sequence—replicates faster than the rest, all the other sequences having the same replication rate. A major discovery made by Eigen is the existence of an error threshold for the mutation rate on the sharp peak landscape: there is a critical mutation rate q_c such that, if $q > q_c$ then the population evolves towards a disordered state, while if $q < q_c$ then the population evolves so as to form a quasispecies, i.e., a population consisting of a positive concentration of the master sequence, along with a cloud of mutants which highly resemble the master sequence.

Eigen’s model is a deterministic model, the population of macromolecules is considered to be infinite and the evolution of the concentrations of the

different genotypes is driven by a system of differential equations. Therefore, when trying to apply the concepts of error threshold and quasispecies to other areas of biology (e.g. population genetics or virology), Eigen's model is not particularly well suited; a model for a finite population, which incorporates stochastic effects, is the most natural mathematical approach to the matter.

Several works have tackled the issue of creating a finite and stochastic version of Eigen's model [1], [5], [6], [10], [11], [12], [13], [14], [15]. Some of these works have recovered the error threshold phenomenon in the case of finite populations: Alves and Fontantari [1] find a relation between the error threshold and the population size by considering a finite version of Eigen's model on the sharp peak landscape. Demetrius, Schuster and Sigmund [5] generalise the error threshold criteria by modelling the evolution of a population via branching processes. Nowak and Schuster [13] also find the error threshold phenomenon in finite populations by making use of a birth and death chain. Some other works have tried to prove the validity of Eigen's model in finite populations by designing algorithms that give similar results to Eigen's theoretical calculations [10], while others have focused on proposing finite population models that converge to Eigen's model in the infinite population limit [6], [12].

The Wright–Fisher model is one of the most classical models in mathematical evolutionary theory, it is also used to understand the evolution of DNA sequences (see [7]). In [3], some counterparts of the results on Eigen's model were derived in the context of the Wright–Fisher model. The Wright–Fisher model describes the evolution of a population of m chromosomes of length ℓ over an alphabet with κ letters. Mutations occur independently at each locus with probability q . The sharp peak landscape is considered: the master sequence replicates at rate $\sigma > 1$, while all the other sequences replicate at rate 1. The following asymptotic regime is studied:

$$\begin{aligned} \ell &\rightarrow +\infty, & m &\rightarrow +\infty, & q &\rightarrow 0, \\ \ell q &\rightarrow a, & \frac{m}{\ell} &\rightarrow \alpha. \end{aligned}$$

In this asymptotic regime the error threshold phenomenon present in Eigen's model is recovered, in the form of a critical curve $\alpha\psi(a) = \ln \kappa$ in the parameter space (a, α) . If $\alpha\psi(a) < \ln \kappa$, then the equilibrium population is totally random, whereas a quasispecies is formed when $\alpha\psi(a) > \ln \kappa$. In the regime where a quasispecies is formed, the concentration of the master sequence in the equilibrium population is also found. The aim of this paper is to continue with the study of the Wright–Fisher model in the above asymptotic regime in

order to find the distribution of the whole quasispecies. It turns out that the resulting distribution is the same as the one found for the Moran model in [4]. Nevertheless, the techniques we use to prove our result are very different from those of [4]. The study of the Moran model relied strongly on monotonicity arguments, and the result was proved inductively. The initial case and the inductive step boiled down to the study of birth and death Markov chains, for which explicit formulas could be found. The Wright–Fisher model is a model with no overlapping generations, for which this approach is no longer suitable. In order to find a more robust approach, we rely on the ideas developed by Freidlin and Wentzell to investigate random perturbations of dynamical systems [9], as well as some techniques already used in [3]. Our setting is essentially the same as the one in [3], the biggest difference being that we work in several dimensions instead of having one dimensional processes. The main challenge is therefore to extend the arguments from [3] to the multidimensional case. This is achieved by replacing the monotonicity arguments employed in [3] by uniform estimates.

We present the main result in the next section. The rest of the paper is devoted to the proof.

2 Main Result

We present the main result of the article here. We start by describing the Wright–Fisher model, we state the result next, and we give a sketch of the proof at the end of the section.

2.1 The Wright–Fisher model

Let \mathcal{A} be a finite alphabet and let κ be its cardinality. Let $\ell, m \geq 1$. Elements of \mathcal{A}^ℓ represent the chromosome of an individual, and we consider a population of m such chromosomes. Two main forces drive the evolution of the population: selection and mutation. The selection mechanism is controlled by a fitness function $A : \mathcal{A}^\ell \rightarrow [0, +\infty[$. We define a selection function $F : \mathcal{A}^\ell \times (\mathcal{A}^\ell)^m \rightarrow [0, 1]$ by setting

$$\forall u \in \mathcal{A}^\ell \quad \forall x \in (\mathcal{A}^\ell)^m \quad F(u, x) = \frac{A(u) \text{card}\{i : 1 \leq i \leq m, x(i) = u\}}{A(x(1)) + \cdots + A(x(m))}.$$

For a given population x , the value $F(u, x)$ is the probability that the individual u is chosen when sampling from x . Throughout the replication process, mutations occur independently on each allele with probability $q \in]0, 1 - 1/\kappa[$. When a mutation occurs, the letter is replaced by a new letter, chosen uniformly at random among the remaining $\kappa - 1$ letters of the alphabet. The mutation mechanism is encoded in a mutation matrix $M(u, v)$, $u, v \in \mathcal{A}^\ell$. The analytical formula for the mutation matrix is as follows:

$$\forall u, v \in \mathcal{A}^\ell \quad M(u, v) = \prod_{j=1}^{\ell} \left((1 - q) 1_{u(j)=v(j)} + \frac{q}{\kappa - 1} 1_{u(j) \neq v(j)} \right).$$

We consider the classical Wright–Fisher model. The transition mechanism from one generation to the next one is divided in two steps. Firstly, we sample with replacement m chromosomes from the current population, according to the selection function F given above. Secondly, each of the sampled chromosomes mutates according to the law given by the mutation matrix. Finally, the whole old generation is replaced with the new one, so generations do not overlap. For $n \geq 0$, we denote by X_n the population at time n , or equivalently, the n -th generation. The Wright–Fisher model is the Markov chain $(X_n)_{n \geq 0}$ with state space $(\mathcal{A}^\ell)^m$, having the following transition matrix:

$$\forall n \in \mathbb{N} \quad \forall x, y \in (\mathcal{A}^\ell)^m$$

$$P(X_{n+1} = y \mid X_n = x) = \prod_{i=1}^m \left(\sum_{u \in \mathcal{A}^\ell} F(u, x) M(u, y(i)) \right).$$

2.2 Main result

We will work only with the sharp peak landscape: there exists a sequence $w^* \in \mathcal{A}^\ell$, called master sequence, whose fitness is $A(w^*) = \sigma > 1$, whereas for all $u \neq w^*$ in \mathcal{A}^ℓ the fitness $A(u)$ is 1. We introduce Hamming classes in the space \mathcal{A}^ℓ . The Hamming distance between two chromosomes $u, v \in \mathcal{A}^\ell$ is defined as follows:

$$d_H(u, v) = \text{card}\{i \in \{1, \dots, \ell\} : u(i) \neq v(i)\}.$$

For $k \in \{1, \dots, \ell\}$ and a population $x \in (\mathcal{A}^\ell)^m$, we denote by $N_k(x)$ the number of sequences in the population x which are at distance k from the master sequence, i.e.,

$$N_k(x) = \text{card}\{i \in \{1, \dots, m\} : d_H(x(i), w^*) = k\}.$$

Let us denote by $I(p, t)$ the rate function governing the large deviations of a binomial law of parameter $p \in [0, 1]$:

$$\forall t \in [0, 1] \quad I(p, t) = t \ln \frac{t}{p} + (1 - t) \ln \frac{1 - t}{1 - p}.$$

We define, for $a \in]0, +\infty[$,

$$\begin{aligned} \forall k \geq 0 \quad \rho_k^* &= (\sigma e^{-a} - 1) \frac{a^k}{k!} \sum_{i \geq 1} \frac{i^k}{\sigma^i}, \\ \rho^*(a) &= \begin{cases} \rho_0^* & \text{if } \sigma e^{-a} > 1 \\ 0 & \text{if } \sigma e^{-a} \leq 1 \end{cases} \\ \psi(a) &= \inf_{l \in \mathbb{N}} \inf \left\{ \sum_{k=1}^{l-1} I\left(\frac{\sigma \rho_k}{(\sigma - 1)\rho_k - 1}, \gamma_k\right) + \gamma_l I\left(e^{-a}, \frac{\rho_{k+1}}{\gamma_k}\right) : \right. \\ &\quad \left. \rho_0 = \rho^*(a), \rho_l = 0, \rho_k, \gamma_k \in [0, 1] \text{ for } 0 \leq k < l \right\}. \end{aligned}$$

Theorem 2.1. *We suppose that*

$$\ell \rightarrow +\infty, \quad m \rightarrow +\infty, \quad q \rightarrow 0,$$

in such a way that

$$\ell q \rightarrow a \in]0, +\infty[, \quad \frac{m}{\ell} \rightarrow \alpha \in [0, +\infty].$$

We have the following dichotomy:

- *if $\alpha \psi(a) < \ln \kappa$, then*

$$\forall k \geq 0 \quad \lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \lim_{n \rightarrow \infty} E\left(\frac{N_k(X_n)}{m}\right) = 0,$$

- *if $\alpha \psi(a) > \ln \kappa$, then*

$$\forall k \geq 0 \quad \lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \lim_{n \rightarrow \infty} E\left(\frac{N_k(X_n)}{m}\right) = \rho_k^*.$$

Moreover, in both cases,

$$\forall k \geq 0 \quad \lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \lim_{n \rightarrow \infty} \text{Var}\left(\frac{N_k(X_n)}{m}\right) = 0.$$

2.3 Sketch of proof

The Wright–Fisher process $(X_n)_{n \geq 0}$ is hard to handle, mainly due to the huge size of the state space and the lack of a natural ordering in it. Instead of directly working with the Wright–Fisher process, we work with the occupancy process $(O_n)_{n \geq 0}$. The occupancy process is a simpler process which derives directly from the original process $(X_n)_{n \geq 0}$, but only keeps the information we are interested in, namely, the number of chromosomes in each of the $\ell + 1$ Hamming classes. The state space of the occupancy process is much simpler than that of the Wright–Fisher process, and it is endowed with a partial ordering. The occupancy process will be the main subject of our study.

We fix next $K \geq 0$ and we focus on finding the concentration of the individuals in the K -th Hamming class. We compare the time that the occupancy process spends having at least one individual in one of the Hamming classes $0, \dots, K$ (persistence time), with the time the process spends having no sequences in any of the classes $0, \dots, K$ (discovery time). Asymptotically, when $\alpha\psi(a) < \ln \kappa$, the persistence time becomes negligible with respect to the discovery time, whereas when $\alpha\psi(a) > \ln \kappa$, it is the discovery time that becomes negligible with respect to the persistence time. This fact, which already proves the first assertion of theorem 2.1, is shown in [3] for the case $K = 0$; the more general case $K \geq 1$ is dealt with in the same way as the case $K = 0$, and the proof does not make any new contributions to the understanding of the model. Therefore, we will admit this fact and focus on the interesting case $\alpha\psi(a) > \ln \kappa$.

We build a coupling to compare the occupancy process with some simpler processes, which will only keep track of the dynamics of the Hamming classes $0, \dots, K$. The simpler processes can be viewed as random perturbations of the same dynamical system. The dynamical system has two fixed points: an unstable one, 0, and a stable one, $\rho^* = (\rho_0^*, \dots, \rho_K^*)$. We use the theory developed by Freidlin and Wentzell [9], as well as some useful estimates from [3], to show that the perturbed processes spend the greatest part of their time very close to the stable fixed point ρ^* , thus showing that the invariant measure of the perturbed processes converge to the Dirac mass in ρ^* .

2.4 The occupancy process

The occupancy process $(O_n)_{n \geq 0}$ will be the starting point of our study. It is obtained from the original Wright–Fisher process $(X_n)_{n \geq 0}$ by using a technique known as lumping (section 4 of [3]). Let $\mathcal{P}_{\ell+1}^m$ be the set of the ordered partitions of the integer m in at most $\ell + 1$ parts:

$$\mathcal{P}_{\ell+1}^m = \{ (o(0), \dots, o(\ell)) \in \mathbb{N}^{\ell+1} : o(0) + \dots + o(\ell) = m \}.$$

A partition $(o(0), \dots, o(\ell))$ is interpreted as an occupancy distribution, which corresponds to a population with $o(l)$ individuals in the Hamming class l , for $0 \leq l \leq \ell$. The occupancy process $(O_n)_{n \geq 0}$ is a Markov chain with values in $\mathcal{P}_{\ell+1}^m$ and transition matrix given by:

$$\forall o, o' \in \mathcal{P}_{\ell+1}^m$$

$$p_O(o, o') = \prod_{0 \leq h \leq \ell} \left(\frac{\sum_{k \in \{0, \dots, \ell\}} o(k) A_H(k) M_H(k, h)}{\sum_{h \in \{0, \dots, \ell\}} o(h) A_H(h)} \right)^{o'(h)},$$

where A_H is the lumped fitness function, defined as follows

$$\forall b \in \{0, \dots, \ell\} \quad A_H(b) = \begin{cases} \sigma & \text{if } b = 0, \\ 1 & \text{if } b \geq 1, \end{cases}$$

and M_H is the lumped mutation matrix: for $b, c \in \{0, \dots, \ell\}$ the coefficient $M_H(b, c)$ is given by

$$\sum_{\substack{0 \leq k \leq \ell-b \\ 0 \leq l \leq b \\ k+l=c-b}} \binom{\ell-b}{k} \binom{b}{l} q^k (1-q)^{\ell-b-k} \left(\frac{q}{\kappa-1} \right)^l \left(1 - \frac{q}{\kappa-1} \right)^{b-l}.$$

The state space $\mathcal{P}_{\ell+1}^m$ of the occupancy process is endowed with a partial order. Let $o, o' \in \mathcal{P}_{\ell+1}^m$, we say that o is lower than or equal to o' , and we write $o \preceq o'$, if

$$\forall l \in \{0, \dots, \ell\} \quad o(0) + \dots + o(l) \leq o'(0) + \dots + o'(l).$$

3 Stochastic bounds

In this section we build simpler processes in order to bound stochastically the occupancy process $(O_n)_{n \geq 0}$. We will couple the simpler processes with the original occupancy process and we will compare their invariant probability measures.

3.1 Lower and upper processes

We begin by constructing a lower process $(O_n^\ell)_{n \geq 0}$ and an upper process $(O_n^{K+1})_{n \geq 0}$ in order to bound stochastically the original occupancy process $(O_n)_{n \geq 0}$. In other words, the lower and upper processes will be built so that for every occupancy distribution $o \in \mathcal{P}_{\ell+1}^m$, if the three processes start from o , then

$$\forall n \geq 0 \quad O_n^\ell \preceq O_n \preceq O_n^{K+1}.$$

The new processes will have simpler dynamics than the original occupancy process.

Let us describe loosely the dynamics of the lower process. As long as there are no master sequences present in the population, the lower process evolves exactly as the original occupancy process. As soon as a master sequence appears, all the chromosomes in the Hamming classes $K+1, \dots, \ell$ are directly sent to the class ℓ . Moreover, as long as the master sequence remains present in the population, all mutations towards the classes $K+1, \dots, \ell$ are also sent to the Hamming class ℓ . The dynamics of the upper process is similar, this time with the Hamming class ℓ replaced by the class $K+1$. The rest of the section is devoted to formalising this construction.

Let Ψ_O be the coupling map defined in section 5.1 of [3]. We modify this map in order to obtain a lower map Ψ_O^ℓ and an upper map Ψ_O^{K+1} . The coupling map Ψ_O takes two arguments, an occupancy distribution $o \in \mathcal{P}_{\ell+1}^m$ and a matrix $r \in \mathcal{R}$, where \mathcal{R} is the set of matrices of size $m \times (\ell+1)$ with coefficients in $[0, 1]$. The Markov chain $(O_n)_{n \geq 0}$ is built with the help of the map Ψ_O and a sequence $(R_n)_{n \geq 1}$ of independent random matrices with values in \mathcal{R} , the entrances of the same random matrix R_n being independent and identically distributed, with uniform law over the interval $[0, 1]$.

Let us define two maps $\pi_\ell, \pi_{K+1} : \mathcal{P}_{\ell+1}^m \rightarrow \mathcal{P}_{\ell+1}^m$ by setting, for every $o \in \mathcal{P}_{\ell+1}^m$,

$$\begin{aligned} \pi_\ell(o) &= (o(0), \dots, o(K), 0, \dots, 0, m - o((0) + \dots + o(K))), \\ \pi_{K+1}(o) &= (o(0), \dots, o(K), m - (o(0) + \dots + o(K)), 0, \dots, 0). \end{aligned}$$

Obviously,

$$\forall o \in \mathcal{P}_{\ell+1}^m \quad \pi_\ell(o) \preceq o \preceq \pi_{K+1}(o).$$

We denote by \mathcal{W}^* the set of occupancy distributions having at least one master sequence, i.e.,

$$\mathcal{W}^* = \{ o \in \mathcal{P}_{\ell+1}^m : o(0) \geq 1 \},$$

and we denote by \mathcal{N} the set of occupancy distributions having no master sequences, i.e.,

$$\mathcal{N} = \{o \in \mathcal{P}_{\ell+1}^m : o(0) = 0\}.$$

Let us define

$$o_{\text{enter}}^\ell = (1, 0, \dots, 0, m-1), \quad o_{\text{exit}}^\ell = (0, \dots, 0, m).$$

The occupancy distributions o_{enter}^ℓ and o_{exit}^ℓ are the absolute minima of the sets \mathcal{W}^* and \mathcal{N} . We define the lower map Ψ_O^ℓ by setting, for $o \in \mathcal{P}_{\ell+1}^m$ and $r \in \mathcal{R}$,

$$\Psi_O^\ell(o, r) = \begin{cases} \Psi_O(o, r) & \text{if } o \in \mathcal{N} \text{ and } \Psi_O(o, r) \notin \mathcal{W}^*, \\ o_{\text{enter}}^\ell & \text{if } o \in \mathcal{N} \text{ and } \Psi_O(o, r) \in \mathcal{W}^*, \\ \pi_\ell(\Psi_O(\pi_\ell(o), r)) & \text{if } o \in \mathcal{W}^* \text{ and } \Psi_O(\pi_\ell(o), r) \notin \mathcal{N}, \\ o_{\text{exit}}^\ell & \text{if } o \in \mathcal{W}^* \text{ and } \Psi_O(\pi_\ell(o), r) \in \mathcal{N}. \end{cases}$$

Likewise, we define the occupancy distributions

$$o_{\text{enter}}^{K+1} = (m, 0, \dots, 0), \quad o_{\text{exit}}^{K+1} = (0, m, 0, \dots, 0),$$

which are the absolute maxima of the sets \mathcal{W}^* and \mathcal{N} . We define an upper map Ψ_O^{K+1} by setting, for $o \in \mathcal{P}_{\ell+1}^m$ and $r \in \mathcal{R}$,

$$\Psi_O^{K+1}(o, r) = \begin{cases} \Psi_O(o, r) & \text{if } o \in \mathcal{N} \text{ and } \Psi_O(o, r) \notin \mathcal{W}^*, \\ o_{\text{enter}}^{K+1} & \text{if } o \in \mathcal{N} \text{ and } \Psi_O(o, r) \in \mathcal{W}^*, \\ \pi_{K+1}(\Psi_O(\pi_{K+1}(o), r)) & \text{if } o \in \mathcal{W}^* \text{ and } \Psi_O(\pi_{K+1}(o), r) \notin \mathcal{N}, \\ o_{\text{exit}}^{K+1} & \text{if } o \in \mathcal{W}^* \text{ and } \Psi_O(\pi_{K+1}(o), r) \in \mathcal{N}. \end{cases}$$

The coupling map Ψ_O is monotone —lemma 5.5 of [3]— i.e., for every pair of occupancy distributions o, o' and for every $r \in \mathcal{R}$,

$$o \preceq o' \implies \Psi_O(o, r) \preceq \Psi_O(o', r).$$

We deduce that the lower map Ψ_O^ℓ is below the coupling map Ψ_O and the upper map Ψ_O^{K+1} is above the coupling map Ψ_O , i.e.,

$$\forall o \in \mathcal{P}_{\ell+1}^m \quad \forall r \in \mathcal{R} \quad \Psi^\ell(o, r) \preceq \Psi_O(o, r) \preceq \Psi_O^{K+1}(o, r).$$

We use the lower and upper maps, along with the i.i.d. sequence of random matrices $(R_n)_{n \geq 0}$, in order to build a lower occupancy process $(O_n^\ell)_{n \geq 0}$ and an upper occupancy process $(O_n^{K+1})_{n \geq 0}$. Let $o \in \mathcal{P}_{\ell+1}^m$ be the starting point of the processes. We set $O_0^\ell = O_0^{K+1} = o$ and

$$\forall n \geq 1 \quad O_n^\ell = \Psi^\ell(O_{n-1}^\ell, R_n), \quad O_n^{K+1} = \Psi^{K+1}(O_{n-1}^{K+1}, R_n).$$

Proposition 3.1. *Suppose that the processes $(O_n)_{n \geq 0}$, $(O_n^\ell)_{n \geq 0}$, $(O_n^{K+1})_{n \geq 0}$ start all from the same occupancy distribution o . We have*

$$\forall n \geq 0 \quad O_n^\ell \preceq O_n \preceq O_n^{K+1}.$$

The proof is similar to the proof of proposition 8.1 in [2].

3.2 Dynamics of the bounding processes

We study now the dynamics of the lower and upper processes in \mathcal{W}^* . Since the calculations are the same for both processes, we take θ to be either $K+1$ or ℓ , and we denote by $(O_n^\theta)_{n \geq 0}$ the corresponding process. For the process $(O_n^\theta)_{n \geq 0}$, the states in the set

$$\mathcal{T}^\theta = \{ o \in \mathcal{P}_{\ell+1}^m : o(0) \geq 1 \text{ and } o(0) + \dots + o(K) + o(\theta) < m \},$$

are transient, and the states in $\mathcal{N} \cup (\mathcal{W}^* \setminus \mathcal{T}^\theta)$ form a recurrence class. Let us take a look at the transition mechanism restricted to $\mathcal{N} \cup (\mathcal{W}^* \setminus \mathcal{T}^\theta)$. Since

$$\mathcal{W}^* \setminus \mathcal{T}^\theta = \{ o \in \mathcal{P}_{\ell+1}^m : o(0) \geq 1 \text{ and } o(0) + \dots + o(K) + o(\theta) = m \},$$

a state in $\mathcal{W}^* \setminus \mathcal{T}^\theta$ is totally determined by the occupancy numbers of the Hamming classes $0, \dots, K$; whenever the process $(O_n^\theta)_{n \geq 0}$ starts from a state in $\mathcal{W}^* \setminus \mathcal{T}^\theta$, the dynamics of $(O_n^\theta(0), \dots, O_n^\theta(K))_{n \geq 0}$ is Markovian until the time of exit from $\mathcal{W}^* \setminus \mathcal{T}^\theta$. Let us define the set

$$\mathbb{D} = \{ z \in \mathbb{N}^{K+1} : z_0 + \dots + z_K \leq m \}.$$

We define the projection $\pi : \mathcal{P}_{\ell+1}^m \rightarrow \mathbb{D}$ by setting, for $o \in \mathcal{P}_{\ell+1}^m$,

$$\pi(o) = (o(0), \dots, o(K)).$$

We denote by $(Z_n^\theta)_{n \geq 0}$ the Markov chain with state space \mathbb{D} and transition matrix given by: for $z, z' \in \mathbb{D}$ and for any $n \geq 0$, let o be the unique element of $\mathcal{P}_{\ell+1}^m \setminus \mathcal{T}^\theta$ such that $\pi(o) = z$,

- if $z_0, z'_0 \geq 1$,

$$P(Z_{n+1}^\theta = z' \mid Z_n^\theta = z) = P(\pi(O_{n+1}^\theta) = z' \mid O_n^\theta = o).$$

- if $z_0 \geq 1$ and $z'_0 = 0$,

$$P(Z_{n+1}^\theta = z_{\text{exit}}^\theta \mid Z_n^\theta = z) = \sum_{z': z'_0=0} P(\pi(O_{n+1}^\theta) = z' \mid O_n^\theta = o),$$

where $z_{\text{exit}}^\ell = (0, \dots, 0)$ and $z_{\text{exit}}^{K+1} = (0, m, 0, \dots, 0)$.

- if $z = z_{\text{exit}}^\theta$,

$$P(Z_{n+1}^\theta = z_{\text{enter}}^\theta \mid Z_n^\theta = z_{\text{exit}}^\theta) = 1,$$

where $z_{\text{enter}}^\ell = (1, 0, \dots, 0)$ and $z_{\text{enter}}^{K+1} = (m, 0, \dots, 0)$.

The remaining non-diagonal coefficients of the transition matrix are null. The diagonal coefficients are chosen so that the matrix is stochastic, i.e., each row adds up to 1. Let us denote by $p^\theta(z, z')$ the above transition matrix and let us compute its value for $z, z' \in \mathbb{D}$ such that $z_0, z'_0 \geq 1$. We introduce some notation first. For $d \geq 1$ and a vector $v \in \mathbb{R}^d$, we denote by $|v|_1$ the L^1 norm of v :

$$|v|_1 = |v_1| + \dots + |v_d|.$$

For $d \geq 1$, a square matrix $M \in \mathbb{R}^{d^2}$, and $i \in \{1, \dots, d\}$, we denote by $M(i, \cdot)$ or $M_{i\cdot}$ the i -th row of M , and by $M(\cdot, i)$ or $M_{\cdot i}$ the i -th column of M . We also denote by $|M|_1$ the L^1 norm of M in \mathbb{R}^{d^2} :

$$|M|_1 = \sum_{i,j=1}^d |M_{ij}|.$$

We say that a vector $s \in \mathbb{D}$ is compatible with another vector $z \in \mathbb{D}$, and we write $s \sim z$, if

$$z_i = 0 \Rightarrow s_i = 0 \quad \text{for } i \in \{0, \dots, K\} \quad \text{and} \quad |z|_1 = m \Rightarrow |s|_1 = m.$$

We say that a matrix $b \in \mathbb{N}^{(K+1)^2}$ is compatible with the vectors $s, z' \in \mathbb{D}$, and we write $b \sim (s, z')$, if

$$\forall i \in \{0, \dots, K\} \quad |b(i, \cdot)|_1 \leq s_i \quad \text{and} \quad |b(\cdot, i)|_1 \leq z'_i.$$

Finally, for $i \in \{0, \dots, K\} \cup \{\theta\}$, we define $M_H(i)$ to be the vector of $[0, 1]^{K+1}$ given by

$$M_H(i) = (M_H(i, 0), \dots, M_H(i, K)).$$

Let $z, z' \in \mathbb{D}$ such that $z_0, z'_0 \geq 1$. We now use the transition mechanism of $(O_n^\theta)_{n \geq 0}$ in order to compute the value of $p^\theta(z, z')$:

$$p^\theta(z, z') = \sum_{s \sim z} \sum_{b \sim (s, z')} p^\theta(z, s, b, z'),$$

where $p^\theta(z, s, b, z')$ is the probability that, given $Z_n^\theta = z$:

- for $i \in \{0, \dots, K\}$, s_i individuals from the class i are selected, and $m - |s|_1$ individuals from the class θ are selected. The probability of this event is

$$\frac{m!}{s_0! \cdots s_K! (m - |s|_1)!} \times \frac{(\sigma z_0)^{s_0} z_1^{s_1} \cdots z_K^{s_K} (m - |z|_1)^{m - |s|_1}}{((\sigma - 1)z_0 + m)^m},$$

- for $i, j \in \{0, \dots, K\}$, b_{ij} individuals from the class i mutate to the class j , and $s_i - |b(i, \cdot)|_1$ individuals from the class i mutate to the class θ . For $i \in \{0, \dots, K\}$, the probability of this event is

$$\frac{s_i!}{b_{i0}! \cdots b_{iK}! (s_i - |b(i, \cdot)|_1)!} \times M_H(i, 0)^{b_{i0}} \cdots M_H(i, K)^{b_{iK}} (1 - |M_H(i)|_1)^{s_i - |b(i, \cdot)|_1},$$

- for $j \in \{0, \dots, K\}$, $z'_j - |b(\cdot, j)|_1$ individuals from the class θ mutate to the class j , and $m - |s|_1 - |z'|_1 + |b|_1$ individuals from the class θ do not mutate to any of the classes $\{0, \dots, K\}$. The probability of this event is

$$\frac{(m - |s|_1)!}{(z'_0 - |b(\cdot, 0)|_1)! \cdots (z'_K - |b(\cdot, K)|_1)! (m - |s|_1 - |z'|_1 + |b|_1)!} \\ \times M_H(\theta, 0)^{z'_0 - |b(\cdot, 0)|_1} \cdots M_H(\theta, K)^{z'_K - |b(\cdot, K)|_1} (1 - |M_H(\theta)|_1)^{m - |s|_1 - |z'|_1 + |b|_1}.$$

Finally,

$$p^\theta(z, s, b, z') = \frac{m!}{s_0! \cdots s_K! (m - |s|_1)!} \times \frac{(\sigma z_0)^{s_0} z_1^{s_1} \cdots z_K^{s_K} (m - |z|_1)^{m - |s|_1}}{((\sigma - 1)z_0 + m)^m} \times \\ \prod_{i=0}^K \frac{s_i!}{b_{i0}! \cdots b_{iK}! (s_i - |b_{i\cdot}|_1)!} \times M_H(i, 0)^{b_{i0}} \cdots M_H(i, K)^{b_{iK}} (1 - |M_H(i)|_1)^{s_i - |b_{i\cdot}|_1} \\ \times \frac{(m - |s|_1)!}{(z'_0 - |b_{\cdot 0}|_1)! \cdots (z'_K - |b_{\cdot K}|_1)! (m - |s|_1 - |z'|_1 + |b|_1)!} \\ \times M_H(\theta, 0)^{z'_0 - |b_{\cdot 0}|_1} \cdots M_H(\theta, K)^{z'_K - |b_{\cdot K}|_1} (1 - |M_H(\theta)|_1)^{m - |s|_1 - |z'|_1 + |b|_1}.$$

3.3 Bounds on the invariant measure

Let us denote by $\mu_O, \mu_O^\ell, \mu_O^{K+1}$ the invariant probability measures of the processes $(O_n)_{n \geq 0}, (O_n^\ell)_{n \geq 0}, (O_n^{K+1})_{n \geq 0}$. Let ν be the image measure of μ_O through the map

$$o \in \mathcal{P}_{\ell+1}^m \mapsto \frac{o(0) + \cdots + o(K)}{m} = \frac{|\pi(o)|_1}{m} \in [0, 1].$$

For every function $g : [0, 1] \mapsto \mathbb{R}$,

$$\int_{[0,1]} g d\nu = \int_{\mathcal{P}_{\ell+1}^m} g\left(\frac{|\pi(o)|_1}{m}\right) d\mu_O = \lim_{n \rightarrow \infty} E\left(g\left(\frac{|\pi(O_n)|_1}{m}\right)\right).$$

Let now $g : [0, 1] \mapsto \mathbb{R}$ be an increasing function such that $g(0) = 0$. Thanks to proposition 3.1, the following inequalities hold: for all $n \geq 0$,

$$g\left(\frac{|\pi(O_n^\ell)|_1}{m}\right) \leq g\left(\frac{|\pi(O_n)|_1}{m}\right) \leq g\left(\frac{|\pi(O_n^{K+1})|_1}{m}\right).$$

Taking the expectation and sending n to ∞ we deduce that

$$\int_{\mathcal{P}_{\ell+1}^m} g\left(\frac{|\pi(o)|_1}{m}\right) d\mu_O^\ell(o) \leq \int_{[0,1]} g d\nu \leq \int_{\mathcal{P}_{\ell+1}^m} g\left(\frac{|\pi(o)|_1}{m}\right) d\mu_O^{K+1}(o).$$

Next, we seek to estimate the above integrals. The strategy is the same for the lower and upper integrals; we set θ to be either $K+1$ or ℓ and we study the invariant probability measure μ_O^θ . We will rely on the following renewal result. Let \mathcal{E} be a finite set and let $(X_n)_{n \geq 0}$ be an ergodic Markov chain with state space \mathcal{E} and invariant probability measure μ . Let \mathcal{W}^* be a subset of \mathcal{E} and let $e \in \mathcal{E}$ be a state outside \mathcal{W}^* . We define

$$\tau^* = \inf\{n \geq 0 : X_n \in \mathcal{W}^*\}, \quad \tau = \inf\{n \geq \tau^* : X_n = e\}.$$

Proposition 3.2. *For every function $f : \mathcal{E} \mapsto \mathbb{R}$, we have*

$$\int_{\mathcal{E}} f d\mu = \frac{E\left(\sum_{n=0}^{\tau-1} f(X_n) \mid X_0 = e\right)}{E(\tau \mid X_0 = e)}.$$

The proof is standard and similar to that of proposition 9.2 of [2]. We apply the renewal result to the process $(O_n^\theta)_{n \geq 0}$ restricted to $\mathcal{N} \cup (\mathcal{W}^* \setminus \mathcal{T}^\theta)$, the set $\mathcal{W}^* \setminus \mathcal{T}^\theta$, the occupancy distribution o_{exit}^θ and the function $o \mapsto g(|\pi(o)|_1/m)$. We set

$$\tau^* = \inf\{n \geq 0 : O_n^\theta \in \mathcal{W}^*\}, \quad \tau = \inf\{n \geq \tau^* : O_n^\theta = o_{\text{exit}}^\theta\}.$$

Applying the renewal theorem we get

$$\int_{\mathcal{P}_{\ell+1}^m} g\left(\frac{|\pi(o)|_1}{m}\right) d\mu_O^\theta(o) = \frac{E\left(\sum_{n=0}^{\tau-1} g\left(\frac{|\pi(O_n^\theta)|_1}{m}\right) \mid O_0^\theta = o_{\text{exit}}^\theta\right)}{E(\tau \mid O_0^\theta = o_{\text{exit}}^\theta)}.$$

Whenever the process $(O_n^\theta)_{n \geq 0}$ is in $\mathcal{W}^* \setminus \mathcal{T}^\theta$, the dynamics of the first $K+1$ Hamming classes, $(\pi(O_n^\theta))_{n \geq 0}$, is that of the Markov chain $(Z_n^\theta)_{n \geq 0}$ defined at the end of the previous section. Let us suppose that $(Z_n^\theta)_{n \geq 0}$ starts from $z_{\text{enter}}^\theta \in \mathbb{D}$, where $z_{\text{enter}}^\theta = (1, 0, \dots, 0)$ and $z_{\text{enter}}^{K+1} = (m, 0, \dots, 0)$. Let τ_0 be the first time that $Z_n^\theta(0)$ becomes null:

$$\tau_0 = \inf\{n \geq 0 : Z_n^\theta(0) = 0\}.$$

Since the process $(O_n^\theta)_{n \geq 0}$ always enters the set $\mathcal{W}^* \setminus \mathcal{T}^\theta$ at the state o_{enter}^θ , the law of τ_0 is the same as the law of $\tau - \tau^*$ for the process $(O_n^\theta)_{n \geq 0}$ starting from o_{exit}^θ . We conclude that the trajectories $(\pi(O_n^\theta))_{\tau^* \leq n \leq \tau}$ and $(Z_n^\theta)_{0 \leq n \leq \tau_0}$ have the same law. Therefore,

$$\begin{aligned} E(\tau - \tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) &= E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta), \\ E\left(\sum_{n=\tau^*}^{\tau-1} g\left(\frac{|\pi(O_n^\theta)|_1}{m}\right) \mid O_0^\theta = o_{\text{exit}}^\theta\right) &= E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n^\theta|_1}{m}\right) \mid Z_0^\theta = z_{\text{enter}}^\theta\right). \end{aligned}$$

Thus, we can rewrite the formula for the invariant probability measure μ_O^θ as follows:

$$\begin{aligned} \int_{\mathcal{P}_{\ell+1}^m} g\left(\frac{|\pi(o)|_1}{m}\right) d\mu_O^\theta(o) &= \frac{E\left(\sum_{n=0}^{\tau^*-1} g\left(\frac{|\pi(O_n^\theta)|_1}{m}\right) \mid O_0^\theta = o_{\text{exit}}^\theta\right)}{E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) + E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta)} \\ &\quad + \frac{E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n^\theta|_1}{m}\right) \mid Z_0^\theta = z_{\text{enter}}^\theta\right)}{E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) + E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta)}. \end{aligned}$$

The objective of the following sections is to estimate each of the terms appearing in the right hand side of this formula.

4 Replicating Markov chains

We study now the Markov chains $(Z_n^\ell)_{n \geq 0}$ and $(Z_n^{K+1})_{n \geq 0}$. The computations are the same for both processes, we take θ to be either $K+1$ or ℓ and we study the Markov chain $(Z_n^\theta)_{n \geq 0}$. We will carry out all of our estimates in the asymptotic regime

$$\ell \rightarrow +\infty, \quad m \rightarrow +\infty, \quad q \rightarrow 0, \quad \ell q \rightarrow a \in]0, +\infty[.$$

We will say that a property holds asymptotically, if it holds for ℓ, m large enough, q small enough and ℓq close enough to a .

4.1 Large deviations for the transition matrix

We define the set $\mathcal{D} \subset \mathbb{R}^{K+1}$ by

$$\mathcal{D} = \left\{ r \in \mathbb{R}^{K+1} : r_0 \geq 0, \dots, r_K \geq 0 \text{ and } r_0 + \dots + r_K \leq 1 \right\}.$$

For $p, t \in \mathcal{D}$, we define the quantity $I_K(p, t)$ as follows:

$$I_K(p, t) = \sum_{k=0}^K t_k \ln \frac{t_k}{p_k} + (1 - |t|_1) \ln \frac{1 - |t|_1}{1 - |p|_1},$$

We make the convention that $a \ln(a/b) = 0$ if $a = b = 0$. The function $I_K(p, \cdot)$ is the rate function governing the large deviations of a multinomial distribution with parameters n and $p_0, \dots, p_K, 1 - |p|_1$. We have the following estimate for the multinomial coefficients:

Lemma 4.1. *For all $n \geq 1$, $N < n$ and $i_0, \dots, i_N \in \{0, \dots, n\}$ such that $s = i_0 + \dots + i_N \leq n$, we have*

$$\left| \ln \frac{n!}{i_0! \dots i_N! (n-s)!} + \sum_{k=0}^N i_k \ln \frac{i_k}{n} + (n-s) \ln \frac{n-s}{n} \right| \leq (N+2) \ln n + 2N + 3.$$

The proof is similar to that of lemma 7.1 of [3].

We define a function $f : \mathcal{D} \rightarrow \mathcal{D}$ by setting

$$\forall r \in \mathcal{D} \quad f(r) = \frac{1}{(\sigma - 1)r_0 + 1} (\sigma r_0, r_1, \dots, r_K).$$

We also define a function $I_\ell : \mathcal{D} \times \mathcal{D} \times [0, 1]^{(K+1)^2} \times \mathcal{D} \rightarrow [0, +\infty]$ by setting, for $r, \xi, t \in \mathcal{D}$ and $\beta \in [0, 1]^{(K+1)^2}$,

$$\begin{aligned} I_\ell(r, \xi, \beta, t) &= I_K(f(r), \xi) + \sum_{k=0}^K \xi_k I_K(M_H(k), \xi_k^{-1} \beta(k, \cdot)) \\ &+ (1 - |\xi|_1) I_K(M_H(\theta), (1 - |\xi|_1)^{-1} (t_0 - |\beta(\cdot, 0)|_1, \dots, t_K - |\beta(\cdot, K)|_1)). \end{aligned}$$

Thanks to the previous identities, for all $z, z', s \in \mathbb{D}$ and $b \in \mathbb{N}^{(K+1)^2}$, we can

express the logarithm of the transition probability $p^\theta(z, s, b, z')$ as follows:

$$\begin{aligned} \ln p^\theta(z, s, b, z') &= -m I_K\left(f\left(\frac{z}{m}\right), \frac{s}{m}\right) - \sum_{k=0}^K s_k I_K(M_H(k), s_k^{-1} b(k, \cdot)) \\ &\quad - (m - |s|_1) I_K\left(M_H(\theta), (m - |s|_1)^{-1} (z'_0 - |b(\cdot, 0)|_1, \dots, z'_K - |b(\cdot, K)|_1)\right) \\ &\quad + \Phi(z, s, b, z') = -m I_\ell\left(\frac{z}{m}, \frac{s}{m}, \frac{b}{m}, \frac{z'}{m}\right) + \Phi(z, s, b, z'). \end{aligned}$$

The error term $\Phi(z, s, b, z')$ satisfies, for all $m \geq 1$,

$$\forall z, z', s \in \mathbb{D} \quad \forall b \in \mathbb{N}^{(K+1)^2} \quad |\Phi(z, s, b, z')| \leq C(K)(\ln m + 1),$$

where $C(K)$ is a constant that depends on K but not on m . In the asymptotic regime, for all $i, j \geq 0$,

$$M_H(i, j) \longrightarrow M_\infty(i, j) = \begin{cases} e^{-a} \frac{a^{j-i}}{(j-i)!} & \text{si } i \leq j, \\ 0 & \text{si } i > j. \end{cases}$$

For $k \in \{0, \dots, K\}$, we set

$$M_\infty(k) = (M_\infty(k, 0), \dots, M_\infty(k, K)).$$

For $t \in \mathcal{D}$, we call $\mathcal{B}(t)$ the subset of $[0, 1]^{(K+1)^2}$ of the upper triangular matrices β such that the sum of the columns of β is equal to the vector t , i.e.,

$$\mathcal{B}(t) = \left\{ \beta \in [0, 1]^{(K+1)^2} : \beta_{ij} = 0 \text{ for } i > j \text{ and } |\beta(\cdot, k)|_1 = t_k \text{ for } 0 \leq k \leq K \right\}.$$

In the asymptotic regime, for $r, \xi, t \in \mathcal{D}$ and $\beta \in [0, 1]^{(K+1)^2}$, we get

$$I_\ell(r, \xi, \beta, t) \longrightarrow \begin{cases} I(r, \xi, \beta, t) & \text{if } \beta \in \mathcal{B}(t), \\ +\infty & \text{otherwise,} \end{cases}$$

where the function $I(r, \xi, \beta, t)$ is given by

$$I(r, \xi, \beta, t) = I_K(f(r), \xi) + \sum_{k=0}^K \xi_k I_K(M_\infty(k), \xi_k^{-1} \beta(k, \cdot)).$$

We define a function $V_1 : \mathcal{D} \times \mathcal{D} \rightarrow [0, \infty]$ by setting, for $r, t \in \mathcal{D}$,

$$V_1(r, t) = \inf \left\{ I(r, \xi, \beta, t) : \xi \in \mathcal{D}, \beta \in \mathcal{B}(t) \right\}.$$

For $r \in \mathbb{R}^{K+1}$, we denote by $\lfloor r \rfloor$ the vector $\lfloor r \rfloor = (\lfloor r_0 \rfloor, \dots, \lfloor r_K \rfloor)$.

Proposition 4.2. *The one step transition probabilities of the Markov chain $(Z_n^\theta)_{n \geq 0}$ verify the large deviations principle governed by V_1 :*

- *For any subset U of \mathcal{D} and for any $\rho \in \mathcal{D}$, we have, for $n \geq 0$,*

$$-\inf \{ V_1(\rho, t) : t \in \overset{\circ}{U} \} \leq \liminf_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln P(Z_{n+1}^\theta \in mU \mid Z_n^\theta = \lfloor m\rho \rfloor).$$

- *For any subsets U, U' of \mathcal{D} , we have, for $n \geq 0$,*

$$\begin{aligned} \limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln \sup_{z \in mU} P(Z_{n+1}^\theta \in mU' \mid Z_n^\theta = z) \\ \leq -\inf \{ V_1(r, t) : r \in \overline{U}, t \in \overline{U'} \}. \end{aligned}$$

Proof. We begin by showing the large deviations upper bound. Let U, U' be two subsets of \mathcal{D} and take $z \in mU$. For $n \geq 0$,

$$\begin{aligned} P(Z_{n+1}^\theta \in mU' \mid Z_n^\theta = z) &= \sum_{z' \in mU' \cap \mathbb{D}} p^\theta(z, z') \\ &= \sum_{z' \in mU' \cap \mathbb{D}} \sum_{s \sim z} \sum_{b \sim (s, z')} p^\theta(z, s, b, z'). \end{aligned}$$

Thanks to the estimates on p^θ , we have, for $m \geq 1$,

$$\begin{aligned} &\sup_{z \in mU} P(Z_{n+1}^\theta \in mU' \mid Z_n^\theta = z) \\ &\leq (m+1)^{C(K)} \max \{ p^\theta(z, s, b, z') : z \in mU, s \sim z, z' \in mU', b \sim (s, z') \} \\ &\leq (m+1)^{C(K)} \exp \left(-m \min \left\{ I_\ell \left(\frac{z}{m}, \frac{s}{m}, \frac{b}{m}, \frac{z'}{m} \right) : \begin{array}{l} z \in mU, z' \in mU' \\ s \sim z, b \sim (s, z') \end{array} \right\} \right), \end{aligned}$$

where $C(K)$ is a constant depending on K but not on m . For each $m \geq 1$, let $z_m, s_m, z'_m \in \mathbb{D}$, $b_m \in \{0, \dots, m\}^{(K+1)^2}$ be four terms that realise the above minimum. We observe next the expression

$$\limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} -I_\ell \left(\frac{z_m}{m}, \frac{s_m}{m}, \frac{b_m}{m}, \frac{z'_m}{m} \right).$$

Since \mathcal{D} and $[0, 1]^{(K+1)^2}$ are compact sets, up to the extraction of a subsequence, we can suppose that when $m \rightarrow \infty$,

$$\frac{z_m}{m} \rightarrow \rho \in \overline{U}, \quad \frac{s_m}{m} \rightarrow \xi \in \mathcal{D}, \quad \frac{b_m}{m} \rightarrow \beta \in [0, 1]^{(K+1)^2}, \quad \frac{z'_m}{m} \rightarrow t \in \overline{U'}.$$

If β is not an upper triangular matrix, or if, for some $j \in \{0, \dots, K\}$, $|\beta(\cdot, j)| \neq t_j$, the limit is $-\infty$. Thus, the only case we need to take care of is when $\beta \in \mathcal{B}(t)$. In this case, we have

$$\limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} -I_\ell\left(\frac{z_m}{m}, \frac{s_m}{m}, \frac{b_m}{m}, \frac{z_m}{m}\right) \leq -I(\rho, \xi, \beta, t).$$

Optimising with respect to ρ, ξ, β, t , we obtain the upper bound of the large deviations principle.

We show next the lower bound. Let $\xi, t \in \mathcal{D}$ and $\beta \in \mathcal{B}(t)$. We have

$$\begin{aligned} P(Z_{n+1}^\theta = \lfloor mt \rfloor \mid Z_n^\theta = \lfloor m\rho \rfloor) &\geq p^\theta(\lfloor m\rho \rfloor, \lfloor m\xi \rfloor, \lfloor m\beta \rfloor, \lfloor mt \rfloor) \\ &\geq (m+1)^{-C(K)} \exp\left(-mI_\ell\left(\frac{\lfloor m\rho \rfloor}{m}, \frac{\lfloor m\xi \rfloor}{m}, \frac{\lfloor m\beta \rfloor}{m}, \frac{\lfloor mt \rfloor}{m}\right)\right). \end{aligned}$$

We take the logarithm and we send m, ℓ to ∞ and q to 0. We obtain then

$$\liminf_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln P(Z_{n+1}^\theta = \lfloor tm \rfloor \mid Z_n^\theta = \lfloor \rho m \rfloor) \geq -I(\rho, \xi, \beta, t).$$

Moreover, if $t \in \overset{\circ}{U}$, for m large enough, $\lfloor tm \rfloor$ belongs to mU . Therefore,

$$\liminf_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln P(Z_{n+1}^\theta \in mU \mid Z_n^\theta = \lfloor m\rho \rfloor) \geq -I(\rho, \xi, \beta, t).$$

We optimise over ξ, β, t and we obtain the large deviations lower bound. \square

A similar proof shows that the l -step transition probabilities of $(Z_n^\theta)_{n \geq 0}$ also satisfy a large deviations principle. For $l \geq 1$, we define a function V_l on $\mathcal{D} \times \mathcal{D}$ as follows:

$$\begin{aligned} V_l(r, t) = \inf \Big\{ &\sum_{k=0}^{l-1} I(\rho_k, \xi_k, \beta_k, \rho_{k+1}) : \\ &\rho_0 = r, \rho_l = t, \rho_k, \xi_k \in \mathcal{D}, \beta_k \in \mathcal{B}(t) \text{ for } 0 \leq k < l \Big\}. \end{aligned}$$

Corollary 4.3. *For $l \geq 1$, the l -step transition probabilities of $(Z_n^\theta)_{n \geq 0}$ satisfy the large deviations principle governed by V_l :*

- For any subset U of \mathcal{D} and for any $\rho \in \mathcal{D}$, we have, for $n \geq 0$,

$$-\inf \{ V_l(\rho, t) : t \in \overset{\circ}{U} \} \leq \liminf_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln P(Z_{n+l}^\theta \in mU \mid Z_n^\theta = \lfloor \rho m \rfloor).$$

- For any subsets U, U' of \mathcal{D} , we have, for $n \geq 0$,

$$\begin{aligned} \limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln \sup_{z \in mU} P(Z_{n+l}^\theta \in mU' \mid Z_n^\theta = z) \\ \leq -\inf \{ V_l(r, t) : r \in \overline{U}, t \in \overline{U'} \}. \end{aligned}$$

4.2 Perturbed dynamical system

We look next for the zeros of the rate function $I(r, \xi, \beta, t)$. We see that $I(r, \xi, \beta, t) = 0$ if and only if $\xi = f(r)$, $\beta \in \mathcal{B}(t)$ and $\beta(k, \cdot)/\xi_k = M_\infty(k)$ for $0 \leq k \leq K$. We define a function $F = (F_0, \dots, F_K) : \mathcal{D} \rightarrow \mathcal{D}$ by setting, for $r \in \mathcal{D}$ and $k \in \{0, \dots, K\}$,

$$F_k(r) = \sum_{i=0}^k f_i(r) e^{-a} \frac{a^{k-i}}{(k-i)!}.$$

Replacing f by its value in the above formula, we can rewrite, for $0 \leq k \leq K$,

$$F_k(r) = \frac{e^{-a}}{(\sigma - 1)r_0 + 1} \left(\frac{a^k}{k!} \sigma r_0 + \sum_{i=1}^k \frac{a^{k-i}}{(k-i)!} r_i \right).$$

The Markov chain $(Z_n^\theta)_{n \geq 0}$ can be seen as a perturbation of the dynamical system associated to the map F :

$$z^0 \in \mathcal{D}, \quad \forall n \geq 1 \quad z^n = F(z^{n-1}).$$

Let ρ^* be the point of \mathcal{D} given by:

$$\forall k \in \{0, \dots, K\} \quad \rho_k^* = (\sigma e^{-a} - 1) \frac{a^k}{k!} \sum_{i \geq 1} \frac{i^k}{\sigma^i}.$$

Proposition 4.4. *We have the following dichotomy:*

- if $\sigma e^{-a} \leq 1$, the function F has a single fixed point, 0, and $(z^n)_{n \geq 0}$ converges to 0.
- if $\sigma e^{-a} > 1$, the function F has two fixed points, 0 and ρ^* . If $z_0^0 = 0$, the sequence $(z^n)_{n \in \mathbb{N}}$ converges to 0, whereas if $z_0^0 > 0$, the sequence $(z^n)_{n \in \mathbb{N}}$ converges to ρ^* .

Proof. For $k \in \{0, \dots, K\}$, the function $F_k(r)$ is a function of r_0, \dots, r_k only; we can inductively solve the system of equations

$$F_k(r) = r_k, \quad 0 \leq k \leq K.$$

For $k = 0$, we have

$$F_0(r) = \frac{\sigma e^{-a} r_0}{(\sigma - 1)r_0 + 1}.$$

The equation $F_0(r) = r_0$ has two solutions: $r_0 = 0$ and $r_0 = \rho_0^*$. For k in $\{1, \dots, K\}$, we have $F_k(r) = r_k$ if and only if

$$r_k = \frac{e^{-a}}{(\sigma - 1)r_0 + 1 - e^{-a}} \left(\frac{a^k}{k!} \sigma r_0 + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} r_i \right).$$

We end up with a recurrence relation. If the initial condition is $r_0 = 0$, the only solution is $r_k = 0$ for all $k \in \{0, \dots, K\}$, whereas if the initial condition is $r_0 = \rho_0^*$, the only solution is $r_k = \rho_k^*$ for all $k \in \{0, \dots, K\}$, this last assertion is shown in section 2.2 of [4].

It remains to show the convergence. We will show the convergence in the case $\sigma e^{-a} > 1$, $z_0^0 > 0$. The other cases are dealt with in a similar fashion, or are even simpler. We will prove the convergence by induction on the coordinates. Since the function

$$F_0(r) = \frac{\sigma e^{-a} r_0}{(\sigma - 1)r_0 + 1}$$

is increasing, concave, and satisfies $F_0(\rho_0^*) = \rho_0^*$, the sequence $(z_0^n)_{n \geq 0}$ is monotone and converges to ρ_0^* . Let $k \in \{1, \dots, K\}$ and let us suppose that the following limit holds:

$$\lim_{n \rightarrow \infty} (z_0^n, \dots, z_{k-1}^n) = (\rho_0^*, \dots, \rho_{k-1}^*).$$

Let $\varepsilon > 0$. We define two functions $\underline{F}, \overline{F} : [0, 1] \rightarrow [0, 1]$ by setting, for

$\rho \in [0, 1]$,

$$\begin{aligned}\underline{F}(\rho) &= \frac{e^{-a}}{(\sigma - 1)(\rho_0^* + \varepsilon) + 1} \left(\frac{a^k}{k!} \sigma(\rho_0^* - \varepsilon) + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} (\rho_i^* - \varepsilon) + \rho \right), \\ \overline{F}(\rho) &= \frac{e^{-a}}{(\sigma - 1)(\rho_0^* - \varepsilon) + 1} \left(\frac{a^k}{k!} \sigma(\rho_0^* + \varepsilon) + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} (\rho_i^* + \varepsilon) + \rho \right).\end{aligned}$$

By the induction hypothesis, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ and $i \in \{0, \dots, k-1\}$, $|z_i^n - \rho_i^*| < \varepsilon$. We have then, for all $n \geq N$ and for all $\rho \in [0, 1]$,

$$\underline{F}(\rho) \leq F_k(z_0^n, \dots, z_{k-1}^n, \rho) \leq \overline{F}(\rho).$$

We define two sequences, $(\underline{z}^n)_{n \geq N}$ and $(\overline{z}^n)_{n \geq N}$, by setting $\underline{z}^N = \overline{z}^N = z_k^N$ and for $n > N$

$$\underline{z}^n = \underline{F}(\underline{z}^{n-1}), \quad \overline{z}^n = \overline{F}(\overline{z}^{n-1}).$$

Thus, for all $n \geq N$, we have $\underline{z}^n \leq z_k^n \leq \overline{z}^n$. Since $\underline{F}(\rho)$ and $\overline{F}(\rho)$ are affine functions, and for ε small enough their main coefficient is strictly smaller than 1, the sequences $(\underline{z}^n)_{n \geq N}$ and $(\overline{z}^n)_{n \geq N}$ converge to the fixed points of the functions \underline{F} et \overline{F} , which are given by:

$$\begin{aligned}\underline{\rho}_k^* &= \frac{e^{-a}}{(\sigma - 1)(\rho_0^* + \varepsilon) + 1 - e^{-a}} \left(\frac{a^k}{k!} \sigma(\rho_0^* - \varepsilon) + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} (\rho_i^* - \varepsilon) \right), \\ \overline{\rho}_k^* &= \frac{e^{-a}}{(\sigma - 1)(\rho_0^* - \varepsilon) + 1 - e^{-a}} \left(\frac{a^k}{k!} \sigma(\rho_0^* + \varepsilon) + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} (\rho_i^* + \varepsilon) \right).\end{aligned}$$

We let ε go to 0 and we see that

$$\lim_{n \rightarrow \infty} z_k^n = \frac{e^{-a}}{(\sigma - 1)\rho_0^* + 1 - e^{-a}} \left(\frac{a^k}{k!} \sigma \rho_0^* + \sum_{i=1}^{k-1} \frac{a^{k-i}}{(k-i)!} \rho_i^* \right) = \rho_k^*,$$

which finishes the inductive step. \square

4.3 Comparison with the master sequence

In section 3, in order to build the bounding occupancy processes, we have fixed an integer $K \geq 0$ and we have kept the relevant information about the dynamics of the occupancy numbers of the Hamming classes $0, \dots, K$. Let us call $(\Theta_n^\ell)_{n \geq 0}$ and $(\Theta_n^1)_{n \geq 0}$ the lower and upper occupancy processes that

are obtained for $K = 0$, and let us call, as before, $(O_n^\ell)_{n \geq 0}$ and $(O_n^{K+1})_{n \geq 0}$ the lower and upper occupancy processes corresponding to $K > 0$. Let us define the following stopping times:

$$\tau(\Theta^\ell) = \inf\{n \geq 0 : \Theta_n^\ell \in \mathcal{N}\}, \quad \tau(O^{K+1}) = \inf\{n \geq 0 : O_n^{K+1} \in \mathcal{N}\}.$$

We have constructed the processes $(\Theta_n^\ell)_{n \geq 0}, (\Theta_n^1)_{n \geq 0}, (O_n^\ell)_{n \geq 0}, (O_n^{K+1})_{n \geq 0}$ in such a way that they are all coupled and the following relations hold: if the four processes start from the same occupancy distribution $o \in \mathcal{W}^*$, then

$$\begin{aligned} \forall n \in \{0, \dots, \tau(\Theta^\ell)\} \quad & \Theta_n^\ell \preceq O_n^\ell \preceq O_n^{K+1} \preceq \Theta_n^{K+1}, \\ \forall n \in \{0, \dots, \tau(O^{K+1})\} \quad & O_n^{K+1} \preceq \Theta_n^{K+1}. \end{aligned}$$

These inequalities are naturally inherited by the Markov chains derived from the occupancy processes; let $(Z_n^\ell)_{n \geq 0}$ and $(Z_n^{K+1})_{n \geq 0}$ be the Markov chains associated to the processes $(O_n^\ell)_{n \geq 0}$ and $(O_n^{K+1})_{n \geq 0}$, as in the end of section 3.2. Likewise, let $(Y_n^\ell)_{n \geq 0}$ and $(Y_n^1)_{n \geq 0}$ be the Markov chains associated to the processes $(\Theta_n^\ell)_{n \geq 0}$ and $(\Theta_n^1)_{n \geq 0}$. The state space of the Markov chains $(Z_n^\ell)_{n \geq 0}, (Z_n^{K+1})_{n \geq 0}$ is the set \mathbb{D} defined in section 3.2, whereas the state space for the Markov chains $(Y_n^\ell)_{n \geq 0}, (Y_n^1)_{n \geq 0}$ is $\{0, \dots, m\}$. Let us define the following stopping times:

$$\tau(Y^\ell) = \inf\{n \geq 0 : Y_n^\ell = 0\}, \quad \tau(Z^{K+1}) = \inf\{n \geq 0 : Z_n^{K+1}(0) = 0\}.$$

Let $z \in \mathbb{D}$ be such that $z_0 \geq 1$, let the Markov chains $(Z_n^\ell)_{n \geq 0}, (Z_n^{K+1})_{n \geq 0}$ start from z , and let z_0 be the starting point of the Markov chains $(Y_n^\ell)_{n \geq 0}, (Y_n^1)_{n \geq 0}$. The inequalities between the occupancy processes translate to the associated Markov chains as follows:

$$\begin{aligned} \forall n \in \{0, \dots, \tau(Y^\ell)\} \quad & Y_n^\ell \leq Z_n^\ell(0) \leq Z_n^{K+1}(0) \leq Y_n^1, \\ \forall n \in \{0, \dots, \tau(Z^{K+1})\} \quad & Z_n^{K+1}(0) \leq Y_n^1. \end{aligned}$$

The occupancy processes $(\Theta_n^\ell)_{n \geq 0}, (\Theta_n^1)_{n \geq 0}$, along with the associated Markov chains $(Y_n^\ell)_{n \geq 0}, (Y_n^1)_{n \geq 0}$, have been studied in detail in [3]. Thanks to the relations just stated, we will be able to make use of many of the estimates derived in [3]. Let θ be $K + 1$ or ℓ and let us call \tilde{V} the cost function associated to the Markov chain $(Y_n^\theta)_{n \geq 0}$. We will make use of the following results from [3]:

Let us define a function $\tilde{F} : [0, 1] \rightarrow [0, 1]$ as follows:

$$\forall r \in [0, 1] \quad \tilde{F}(r) = e^{-a} \frac{\sigma r}{(\sigma - 1)r + 1}.$$

Lemma 4.5. *Suppose that $\sigma e^{-a} > 1$. For $s, t \in [0, 1]$, we have $\tilde{V}(s, t) = 0$ if and only if*

- *either $s = t = 0$,*
- *or there exists $l \geq 1$ such that $t = \tilde{F}^l(s)$,*
- *or $s \neq 0, t = \rho^*$.*

Let $\tau(Y^\theta)$ be the first time that the Markov chain $(Y_n^\theta)_{n \geq 0}$ becomes null:

$$\tau(Y^\theta) = \inf \{ n \geq 0 : Y_n^\theta = 0 \}.$$

Proposition 4.6. *Let $a \in]0, +\infty[$ and let $i \in \{1, \dots, m\}$. The expected value of $\tau(Y^\theta)$ starting from i satisfies*

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln E(\tau(Y^\theta) | Y_0^\theta = i) = \tilde{V}(\rho^*, 0).$$

4.4 Concentration near ρ^*

We show next that, when $\sigma e^{-a} > 1$, asymptotically, the Markov chain $(Z_n^\theta)_{n \geq 0}$ concentrates in a neighbourhood of ρ^* . Let us loosely describe the strategy we will follow. The Markov chain $(Z_n^\theta)_{n \geq 0}$ is a perturbation of the dynamical system associated to the map F . The map F has two fixed points: 0 and ρ^* . The fixed point 0 is unstable, while ρ^* is a stable fixed point. The proof relies mainly on two different kind of estimates. We estimate first the typical time the process $(Z_n^\theta)_{n \geq 0}$ needs to leave a neighbourhood of the region $\{z \in \mathbb{D} : z_0 = 0\}$; since the instability at 0 concerns principally the dynamics of the master sequence, we will be able to make use of the estimates developed in [3] by means of the inequalities stated in section 4.3. We estimate then the time the process $(Z_n^\theta)_{n \geq 0}$ spends outside a neighbourhood of the region $\{z \in \mathbb{D} : z_0 = 0\}$ and ρ^* . Since $(Z_n^\theta)_{n \geq 0}$ tends to follow the discrete trajectories given by the dynamical system associated to F , it cannot stay a long time outside such a neighbourhood. This fact will be proved with the help of the large deviations principle stated in the previous section. This estimate will help us to bound the number of excursions outside a neighbourhood of ρ^* , as well as the length of these excursions. We formalise these ideas in the rest of the section. In order to simplify the notation, from now

on we omit the superscript θ and we denote by P_z and E_z the probabilities and expectations for the Markov chain $(Z_n)_{n \geq 0}$ starting from $z \in \mathbb{D}$.

Let us define

$$D_\delta = \{ r \in \mathcal{D} : 0 < r_0 < \delta \}.$$

Lemma 4.7. *For all $\delta > 0$, there exists $c > 0$, depending on δ , such that, asymptotically, for all $z \in \mathbb{D}$ such that $z_0 \geq 1$, we have*

$$P_z(Z_1(0) > 0, \dots, Z_{\lfloor c \ln m \rfloor - 1}(0) > 0, Z_{\lfloor c \ln m \rfloor} \in m(\mathcal{D} \setminus \overline{D_\delta})) \geq \frac{1}{m^{c \ln m}}.$$

Proof. Let $(Y_n^\ell)_{n \geq 0}$ be the Markov chain defined in section 4.3. Let $\tau(Y^\ell)$ be the first time that the process $(Y_n^\ell)_{n \geq 0}$ becomes null:

$$\tau(Y^\ell) = \inf \{ n \geq 0 : Y_n = 0 \}.$$

By the remarks in section 4.3 we can see that

$$\begin{aligned} & P_z(Z_1(0) > 0, \dots, Z_{\lfloor c \ln m \rfloor - 1}(0) > 0, Z_{\lfloor c \ln m \rfloor} \in m(\mathcal{D} \setminus \overline{D_\delta})) \geq \\ & P_z(Z_1(0) > 0, \dots, Z_{\lfloor c \ln m \rfloor - 1}(0) > 0, Z_{\lfloor c \ln m \rfloor} \in m(\mathcal{D} \setminus \overline{D_\delta}), \tau(Y^\ell) > \lfloor c \ln m \rfloor) \\ & \geq P_1(Y_1^\ell > 0, \dots, Y_{\lfloor c \ln m \rfloor - 1}^\ell > 0, Y_{\lfloor c \ln m \rfloor}^\ell > m(\rho_0^* - \delta)). \end{aligned}$$

As shown in lemma 7.7 of [3], this last probability is bounded from below by $1/m^{c \ln m}$, which gives the desired result. \square

We estimate next the length of a typical excursion of $(Z_n)_{n \geq 0}$ outside a neighbourhood of $\{z \in \mathbb{D} : z_0 = 0\}$ and $m\rho^*$. For $\rho \in \mathcal{D}$ and $\delta > 0$, we define the δ -neighbourhood of ρ by

$$U(\rho, \delta) = \{ r \in \mathcal{D} : |r - \rho| < \delta \}.$$

Lemma 4.8. *For all $\delta > 0$, there exist $h \geq 1$ and $c > 0$, depending on δ , such that, asymptotically, for all $r \in \mathcal{D}$ such that $r_0 \geq \delta$, we have*

$$P_{\lfloor mr \rfloor}(Z_1(0) > 0, \dots, Z_{h-1}(0) > 0, Z_h \in mU(\rho^*, \delta)) \geq 1 - \exp(-cm).$$

Proof. Let $\delta > 0$ and let us define the set

$$\mathcal{K} = \{ r \in \mathcal{D} : r_0 \geq \delta \}.$$

For each $r \in \mathcal{K}$ there exists an integer $h_r \geq 0$ such that $F^{h_r}(r) \in U(\rho^*, \delta/4)$. By continuity of the map F , for each $r \in \mathcal{K}$ there exist also positive numbers $\delta_0^r, \dots, \delta_{h_r}^r$ such that $\delta_0^r, \dots, \delta_{h_r}^r < \delta/2$ and

$$\forall k \in \{0, \dots, h_r\} \quad F(U(F^{k-1}(r), \delta_{k-1}^r)) \subset U(F^k(r), \delta_k^r/2).$$

The family $\{U(r, \delta_0^r) : r \in \mathcal{K}\}$ is an open cover of the set \mathcal{K} ; since \mathcal{K} is a compact set, we can extract a finite subcover, i.e., there exist $N \in \mathbb{N}$ and $r_1, \dots, r_N \in \mathcal{K}$ such that

$$\mathcal{K} \subset U_0 = \bigcup_{n=1}^N U(r_n, \delta_0^{r_n}).$$

Let us set $h = \max\{h_{r_i} : 1 \leq i \leq N\}$, For $n \in \{1, \dots, N\}$ we take $\delta_{h_{r_n}+1}^{r_n}, \dots, \delta_h^{r_n}$ to be positive numbers such that, as before,

$$\forall k \in \{h_{r_n} + 1, \dots, h\} \quad F(U(F^{k-1}(r_n), \delta_{k-1}^{r_n})) \subset U(F^k(r_n), \delta_k^{r_n}/2).$$

Let us define

$$\forall k \in \{1, \dots, h-1\} \quad U_k = \bigcup_{n=1}^N U(F^k(r_n), \delta_k^{r_n}).$$

We have then, for any $r \in \mathcal{K}$,

$$P_{\lfloor mr \rfloor}(Z_1(0) > 0, \dots, Z_{h-1}(0) > 0, Z_h \in mU(\rho^*, \delta)) \geq P_{\lfloor mr \rfloor}(\forall k \in \{1, \dots, h\} \quad Z_k \in mU_k).$$

Passing to the complementary event,

$$\begin{aligned} & P_{\lfloor mr \rfloor}(\exists k \in \{1, \dots, h-1\} \quad Z_k(0) = 0 \text{ or } Z_h \notin mU(\rho^*, \delta)) \\ & \leq P_{\lfloor mr \rfloor}(\exists k \in \{1, \dots, h\} \quad Z_k \notin mU_k) \\ & \leq \sum_{1 \leq k \leq h} P_{\lfloor mr \rfloor}(Z_1 \in mU_1, \dots, Z_{k-1} \in mU_{k-1}, Z_k \notin mU_k) \\ & \leq \sum_{1 \leq k \leq h} \sum_{z \in mU_{k-1}} P_{\lfloor mr \rfloor}(Z_{k-1} = z, Z_k \notin mU_k) \\ & \leq \sum_{1 \leq k \leq h} \sum_{z \in mU_{k-1}} P_z(Z_k \notin mU_k) P_{\lfloor mr \rfloor}(Z_{k-1} = z) \\ & \leq \sum_{1 \leq k \leq h} \max \{ P_z(Z_1 \notin mU_k) : z \in mU_{k-1} \}. \end{aligned}$$

We use now the large deviations upper bound stated in proposition 4.2. We have, for all $k \in \{1, \dots, h\}$,

$$\limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln \max_{z \in mU_{k-1}} P_z(Z_1 \notin mU_k) \leq \\ - \inf \{ I(r, \xi, \beta, t) : r \in \overline{U_{k-1}}, \xi \in \mathcal{D}, \beta \in \mathcal{B}(t), t \notin U_k \}.$$

For all $r \in \overline{U_{k-1}}$, we have $F(r) \in U_k$, the previous infimum is thus strictly positive. Since h is fixed, we conclude that

$$\limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln P_{\lfloor mr \rfloor}(\exists k \in \{1, \dots, h-1\} \ Z_k = 0 \text{ or } Z_h \notin mU(\rho^*, \delta)) < 0,$$

which finishes the proof of the lemma. \square

Corollary 4.9. *Let $\delta > 0$. There exist $h \geq 1$, $c \geq 0$, depending on δ , such that, asymptotically, for all $r \in \mathcal{D} \setminus (\overline{D}_\delta \cup U(\rho^*, \delta))$ and for all $n \geq 0$, we have*

$$P_{\lfloor mr \rfloor}(Z_t \in \mathcal{D} \setminus (\overline{D}_\delta \cup U(\rho^*, \delta)) \text{ for } 0 \leq t \leq n) \leq \exp\left(-cm \left\lfloor \frac{n}{h} \right\rfloor\right).$$

The proof is carried out by dividing the interval $\{0, \dots, n\}$ in subintervals of length h and using the estimate of lemma 4.8 on each of the subintervals. We will not write the details, which can be found in the proof of corollary 7.10 of [3].

Proposition 4.10. *Let $g : [0, 1] \rightarrow [0, 1]$ be an increasing and continuous function, such that $g(0) = 0$. For all $z^0 \in \mathbb{D}$ such that $z_0^0 \geq 1$, we have*

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n|_1}{m} \middle| Z_0 = z^0\right)\right)}{E(\tau_0 \mid Z_0 = z^0)} = g(|\rho^*|_1).$$

Proof. Let $\varepsilon > 0$. The function g being continuous, there exists $\delta > 0$ such that

$$\forall \rho \in U(\rho^*, 2\delta) \quad |g(|\rho|_1) - g(|\rho^*|_1)| < \varepsilon.$$

We define next a sequence of stopping times in order to control the excursions of the Markov chain $(Z_n)_{n \geq 0}$ outside $U(\rho^*, \delta)$. We take $T_0 = 0$ and

$$\begin{aligned} T_1^* &= \inf \left\{ n \geq 0 : \frac{Z_n}{m} \in U(\rho^*, \delta) \right\} & T_1 &= \inf \left\{ n \geq T_1^* : \frac{Z_n}{m} \notin U(\rho^*, 2\delta) \right\} \\ \vdots & & \vdots & \\ T_k^* &= \inf \left\{ n \geq T_{k-1} : \frac{Z_n}{m} \in U(\rho^*, \delta) \right\} & T_k &= \inf \left\{ n \geq T_k^* : \frac{Z_n}{m} \notin U(\rho^*, 2\delta) \right\} \\ \vdots & & \vdots & \end{aligned}$$

We have then

$$\begin{aligned} \sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n|_1}{m}\right) - g(|\rho^*|_1)\tau_0 &= \sum_{k \geq 1} \sum_{n=T_{k-1}^* \wedge \tau_0}^{T_k^* \wedge \tau_0-1} \left(g\left(\frac{|Z_n|_1}{m}\right) - g(|\rho^*|_1) \right) \\ &\quad + \sum_{k \geq 1} \sum_{n=T_k^* \wedge \tau_0}^{T_k \wedge \tau_0-1} \left(g\left(\frac{|Z_n|_1}{m}\right) - g(|\rho^*|_1) \right). \end{aligned}$$

Taking the absolute value, we obtain

$$\left| \sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n|_1}{m}\right) - g(|\rho^*|_1)\tau_0 \right| \leq 2g(1) \sum_{k \geq 1} (T_k^* \wedge \tau_0 - T_{k-1} \wedge \tau_0) + \varepsilon \tau_0.$$

We need to control the sum on the right hand side. Let us define, for $n \geq 0$,

$$N(n) = \max \{ k \geq 1 : T_{k-1} < n \}.$$

We can now rewrite the sum as follows:

$$\sum_{k \geq 1} (T_k^* \wedge \tau_0 - T_{k-1} \wedge \tau_0) = \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1}).$$

Let $\eta > 0$ and let us take t_m^η as in proposition 7.11 of [3]:

$$t_m^\eta = \exp \left(m(\tilde{V}(\rho_0^*, 0) + \eta) \right),$$

where \tilde{V} is the cost function governing the dynamics of the master sequence, as defined in section 4.3. We decompose the previous sum as follows:

$$\sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1}) \leq 1_{\tau_0 > t_m^\eta} \tau_0 + 1_{\tau_0 \leq t_m^\eta} \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1}).$$

Let $z^0 \in \mathbb{D}$ such that $z_0^0 \geq 1$. Since the estimates are the same for every starting point, we do not write the starting point in what follows: the probabilities and expectation are all taken with respect to the initial condition $Z_0 = z^0$, unless otherwise stated. We take the expectation in the previous inequalities and we obtain

$$\left| E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n|_1}{m}\right)\right) - g(|\rho^*|_1)E(\tau_0) \right| \leq \\ 2g(1)E(1_{\tau_0 > t_m^\eta} \tau_0) + 2g(1)E\left(1_{\tau_0 \leq t_m^\eta} \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})\right) + \varepsilon E(\tau_0).$$

Thanks to the estimates developed in section 7.3 of [3], we know that

$$\lim_{m \rightarrow \infty} E(1_{\tau_0 > t_m^\eta} \tau_0) = 0.$$

Proceeding as in lemma 7.13 of [3], we can obtain the following bound on N :

Lemma 4.11. *There exists $c > 0$, depending on δ , such that, asymptotically,*

$$\forall k, n \geq 0 \quad P(N(n) > k) \leq \frac{n^k}{k!} \exp(-cmk).$$

We estimate next the term

$$E\left(1_{\tau_0 \leq t_m^\eta} \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})\right).$$

We have, by the Cauchy–Schwarz inequality,

$$\begin{aligned} E\left(1_{\tau_0 \leq t_m^\eta} \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})\right) &= \sum_{k \geq 1} E(1_{\tau_0 \leq t_m^\eta} 1_{k \leq N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})) \\ &\leq \sum_{k \geq 1} P(\tau_0 \leq t_m^\eta, N(\tau_0) \geq k)^{1/2} E(1_{k \leq N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})^2)^{1/2} \\ &\leq \sum_{k \geq 1} P(N(t_m^\eta) \geq k)^{1/2} E(1_{k \leq N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})^2)^{1/2}. \end{aligned}$$

If $1 \leq k \leq N(\tau_0)$, then $T_{k-1} < \tau_0$ and $Z_{T_{k-1}}(0) > 0$. Thanks to the Markov

property,

$$\begin{aligned}
& E(1_{k \leq N(\tau_0)}(T_k^* \wedge \tau_0 - T_{k-1})^2) \\
&= \sum_{z \in \mathbb{D}: z_0 \geq 1} E(1_{k \leq N(\tau_0)}(T_k^* \wedge \tau_0 - T_{k-1})^2 \mid Z_{T_{k-1}} = z) \times P(Z_{T_{k-1}} = z) \\
&\leq \sum_{z \in \mathbb{D}: z_0 \geq 1} E_z((T_1^* \wedge \tau_0)^2) P(Z_{T_{k-1}} = z).
\end{aligned}$$

We next seek an upper bound on the random time $T_1^* \wedge \tau_0$, whenever the Markov chain starts from $z \in \mathbb{D}$ with $z_0 \geq 1$.

Lemma 4.12. *For all $\delta > 0$, there exist $h \geq 1$, $c > 0$, depending on δ , such that, asymptotically, for $z \in \mathbb{D}$ such that $z_0 \geq 1$,*

$$P_z(Z_{\lfloor c \ln m \rfloor + h} \in mU(\rho^*, \delta)) \geq \frac{1}{2m^{c \ln m}}.$$

Proof. Thanks to lemma 4.7, there exists $c > 0$, such that, asymptotically, for all $z \in \mathbb{D}$ such that $z_0 \geq 1$,

$$P_z(Z_{\lfloor c \ln m \rfloor}(0) > \delta m) \geq \frac{1}{m^{c \ln m}}.$$

Likewise, thanks to lemma 4.8, there exist $h \geq 1$ and $c' > 0$, such that, asymptotically, for all $z' \in \mathbb{D}$ such that $z'_0 \geq \delta m$,

$$P_{z'}(Z_h \in mU(\rho^*, \delta)) \geq 1 - \exp(-c'm).$$

Thus,

$$\begin{aligned}
& P_z(Z_{\lfloor c \ln m \rfloor + h} \in mU(\rho^*, \delta)) \\
&\geq P_z(Z_{\lfloor c \ln m \rfloor}(0) \geq \delta m, Z_{\lfloor c \ln m \rfloor + h} \in mU(\rho^*, \delta)) = \\
&\sum_{z': z'_0 \geq \delta m} P_z(Z_{\lfloor c \ln m \rfloor} = z') P_{z'}(Z_h \in mU(\rho^*, \delta)) \geq \frac{1}{m^{c \ln m}} (1 - \exp(-c'm)),
\end{aligned}$$

and the result holds. \square

Corollary 4.13. *For all $\delta > 0$, there exist $h \geq 1$, $c > 0$, depending on δ , such that, asymptotically, for all $z \in \mathbb{D}$ such that $z_0 \geq 1$,*

$$\forall n \geq 0 \quad P_z(T_1^* \wedge \tau_0 \geq n(\lfloor c \ln m \rfloor + h)) \leq \left(1 - \frac{1}{2m^{c \ln m}}\right)^n.$$

Again, the proof is done by dividing the interval $\{0, \dots, n(\lfloor c \ln m \rfloor + h)\}$ in subintervals of length $\lfloor c \ln m \rfloor + h$ and using the estimates of lemma 4.12 on each subinterval, as in corollary 7.10 of [3]. Thanks to corollary 4.13, asymptotically, for all $z \in \mathbb{D}_K$ such that $z_0 \geq 1$,

$$E_z((T_1^* \wedge \tau_0)^2) = \sum_{k \geq 1} P_z(T_1^* \wedge \tau_0 \geq \sqrt{k}) \leq \sum_{k \geq 1} \left(1 - \frac{1}{2m^{c \ln m}}\right)^{\left\lfloor \frac{\sqrt{k}}{\lfloor c \ln m \rfloor + h} \right\rfloor}.$$

Let us set

$$\alpha = 1 - \frac{1}{2m^{c \ln m}}, \quad t = \lfloor c \ln m \rfloor + h.$$

We have:

$$\sum_{k \geq 1} \alpha^{\lfloor \sqrt{k}/t \rfloor} \leq \sum_{k \geq 1} \alpha^{\sqrt{k}/t-1} \leq \int_0^\infty \alpha^{\sqrt{x}/t-1} dx = \frac{2t^2}{\alpha(\ln \alpha)^2}.$$

Therefore, asymptotically, for all $z \in \mathbb{D}_K$ such that $z_0 \geq 1$,

$$E_z((T_1^* \wedge \tau_0)^2) \leq m^{3c \ln m}.$$

Thus, for all $k \geq 1$,

$$E(1_{k \leq N(\tau_0)}(T_k^* \wedge \tau_0 - T_{k-1})^2) \leq m^{3c \ln m}.$$

Together with lemma 4.11, this implies that

$$\begin{aligned} E\left(1_{\tau_0 \leq t_m^\eta} \sum_{k=0}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})\right) &\leq \sum_{k \geq 1} P(N(t_m^\eta) > k)^{1/2} (m^{3c \ln m})^{1/2} \\ &\leq m^{3c \ln m} \left(t_m^\eta \exp(-cm/3) + \sum_{k \geq t_m^\eta \exp(-cm/3)} \left(\frac{(t_m^\eta)^k}{k!} \exp(-cmk) \right)^{1/2} \right) \\ &\leq m^{3c \ln m} \left(t_m^\eta \exp(-cm/3) + \sum_{k \geq 0} \exp\left(\frac{k}{2} - cm\frac{k}{3}\right) \right). \end{aligned}$$

The last inequality holds since $k! \geq (k/e)^k$. We choose η such that $0 < \eta < c/3$. Thanks to the preceding inequality,

$$\begin{aligned} \limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a}} \frac{1}{m} \ln E\left(1_{\tau_0 \leq t_m^\eta} \sum_{k=1}^{N(\tau_0)} (T_k^* \wedge \tau_0 - T_{k-1})\right) \\ \leq \max\left(\tilde{V}(\rho_0^*, 0) + \eta - \frac{c}{3}\right) < \tilde{V}(\rho_0^*, 0). \end{aligned}$$

Theses estimates, along with the result of proposition 4.6, imply that

$$\left| E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n|_1}{m}\right)\right) - g(|\rho^*|_1)E(\tau_0) \right| \leq 3\varepsilon E(\tau_0),$$

which concludes the proof of proposition 4.10. \square

5 Synthesis

The first statement of theorem 2.1 is proved in [3] for the case of the master sequence, $K = 0$. The proof for the case $K \geq 1$ does not involve any new arguments or ideas for a better understanding of the model; it is a straightforward generalisation of the proof for the case $K = 0$. Thus we deal only with the second statement of theorem 2.1. Let us suppose that $\alpha\psi(a) > \ln \kappa$. As shown in [3], the following estimates hold: $\forall a \in]0, +\infty[$, $\forall \alpha \in [0, +\infty]$,

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \frac{1}{m} \ln E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta) = \tilde{V}(\rho^*, 0),$$

$$\limsup_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \frac{1}{\ell} \ln E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) \leq \ln \kappa.$$

Thus, since we are studying the case $\alpha\psi(a) > \ln \kappa$,

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \frac{E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta)}{E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta)} = +\infty.$$

On one hand, g being a bounded function, the above identity readily implies that

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \frac{E\left(\sum_{n=0}^{\tau^*-1} g\left(\frac{|\pi(O_n^\theta)|_1}{m}\right) \mid O_0^\theta = o_{\text{exit}}^\theta\right)}{E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) + E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta)} = 0.$$

On the other hand, using proposition 4.10, we see that

$$\lim_{\substack{\ell, m \rightarrow \infty, q \rightarrow 0 \\ \ell q \rightarrow a, \frac{m}{\ell} \rightarrow \alpha}} \frac{E\left(\sum_{n=0}^{\tau_0-1} g\left(\frac{|Z_n^\theta|_1}{m}\right) \mid Z_0^\theta = z_{\text{enter}}^\theta\right)}{E(\tau^* \mid O_0^\theta = o_{\text{exit}}^\theta) + E(\tau_0 \mid Z_0^\theta = z_{\text{enter}}^\theta)} = g(\rho_0^* + \dots + \rho_K^*).$$

Reporting back in the formula at the very end of section 3.3, we conclude the proof of theorem 2.1.

References

- [1] Domingos Alves and Jose Fernando Fontanari. Error threshold in finite populations. *Phys. Rev. E*, 57:7008–7013, 1998.
- [2] Raphaël Cerf. Critical population and error threshold on the sharp peak landscape for a Moran model. *preprint*, 2012.
- [3] Raphaël Cerf. Critical population and error threshold on the sharp peak landscape for the Wright–Fisher model. *preprint*, 2012.
- [4] Raphaël Cerf and Joseba Dalmau. The distribution of the quasispecies for a Moran model on the sharp peak landscape. *preprint*, 2013.
- [5] Lloyd Demetrius, Peter Schuster, and Karl Sigmund. Polynucleotide evolution and branching processes. *Bulletin of Mathematical Biology*, 47(2):239 – 262, 1985.
- [6] Narendra M. Dixit, Piyush Srivastava, and Nisheeth K. Vishnoi. A finite population model of molecular evolution: theory and computation. *J. Comput. Biol.*, 19(10):1176–1202, 2012.
- [7] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.
- [8] Manfred Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.
- [9] Mark I. Freidlin and Alexander D. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, third edition, 2012. Translated from the 1979 Russian original by Joseph Szücs.
- [10] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22:403–434, 1976.

- [11] John McCaskill. A stochastic theory of macromolecular evolution. *Biological Cybernetics*, 50:63–73, 1984.
- [12] Fabio Musso. A stochastic version of the Eigen model. *Bull. Math. Biol.*, 73(1):151–180, 2011.
- [13] Martin A. Nowak and Peter Schuster. Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller’s ratchet. *Journal of theoretical Biology*, 137 (4):375–395, 1989.
- [14] Jeong-Man Park, Enrique Muñoz, and Michael W. Deem. Quasispecies theory for finite populations. *Phys. Rev. E*, 81:011902, 2010.
- [15] David B. Saakian, Michael W. Deem, and Chin-Kun Hu. Finite population size effects in quasispecies models with single-peak fitness landscape. *Europhysics Letters*, 98(1):18001, 2012.